Assessing Causality Assessment Methods

by
David A. Lane, Ph.D., and
Tom A. Hutchinson, M.B.[1]

University of Minnesota
Technical Report No. 460
February 1986

Direct correspondence to: Professor David A. Lane, School of Statistics, 270 Vincent Hall, University of Minnesota, Minneapolis, MN 55455 (612-373-3035).

---

[1]This research has been supported by grants from the University of Minnesota Graduate School, the Drug Information Association, the American Medical Association Education and Research Foundation and Ciba-Geigy Canada. We wish to acknowledge the helpful comments and criticisms we have received during the preparation of this paper from our colleagues and collaborators Drs. Judith Jones, Claudio Naranjo, Michael Kramer, Tom Rector, Bill Sudderth and Seymour Geisser

Abstract

Six criteria that methods for assessing the causality of adverse drug reactions should satisfy are developed and applied to evaluate ten current causality assessment methods. The main conclusion is that all the methods fail to satisfy at least some of the criteria.

Key words: adverse drug reactions, causality assessment, standardized assessment schemes

ASSESSING CAUSALITY ASSESSMENT METHODS

1. Introduction

A drug is administered to a patient, and some time later the patient experiences an adverse clinical event. Did the drug cause the event to occur, or would the event have happened as and when it did had the drug not been administered? Answering this question may be exceedingly complex, and it is nearly always fraught with uncertainty. It is complex, because many factors concerning the patient, the suspected drug, other drugs, and non-drug exposures may contribute to the occurrence of the event. It is uncertain, because the information concerning these factors (or even the event itself) is rarely complete and accessible [10, 17]. Consequently, various methods have been developed during the past ten years to assess the causal links between drugs and adverse events [3, 4, 6, 7, 10, 11, 14, 18, 19, 20, 25, 27, 31].

Although these causality assessment methods differ from one another in many respects, they share a common basic structure. First, they pose a series of questions that elicit pertinent details about a particular case involving a suspected adverse drug reaction. These questions are designed to determine whether the timing and clinical characteristics of the adverse event are consistent with its being an adverse reaction to the drug; whether alternative etiological candidates exist; whether the event, if reversible, abates when the drug is discontinued (dechallenge) and reappears if the drug is readministered to the patient (rechallenge); in addition, most of the methods ask whether the event is generally recognized as a possible adverse reaction to the drug. Second, the methods give procedures for converting the answers to these questions into a measure of the probability that the adverse event under consideration was caused by any of the drugs to which the patient had been exposed.

It is important to clarify what is meant by the probability that a drug D caused an adverse event E, especially since there is some confusion on this point in the causality assessment literature. Most authors have regarded the probability of drug causation as an objective attribute of the particular drug-event connection, that can be determined from publicly accessible and unambiguous evidence that is elicited in response to "operational" questions. According to this view, there is one correct answer to the question of drug causation, and logical processing of the available information will yield that answer (see, for example, [13,14,18]).

We do not believe that this point of view holds up under careful analysis. Rather, we maintain that the probability of drug causation is a subjective measure of the assessor's uncertainty, based on incomplete and perhaps conflicting information. In other words, the available evidence does not permit a "correct" answer. In reality, D either did or did not cause E, in the sense that E either would or would not have occurred as and when it did, had D not been administered. The difficulty is that one cannot, on the strength of the information available, determine with certainty which of these alternatives is correct.

If nothing depended on it, the extent of the uncertainty about whether D caused E would be a matter of no concern. But in many situations one would act differently if he knew that D caused E than if he knew that it did not: for example, a clinician might immediately cease therapy with D if he knew it caused E, while if he were certain that it did not, he would continue administering D and prescribe treatment for E; or, the experimenters might discontinue a clinical trial on a new drug D if they knew for sure that it had caused a subject's death, while they would continue the trial if the death were unrelated to D. In such cases, what one decides to do depends in part on how likely he believes it is that D caused E, and so it is important to measure the extent or degree of this belief as a step in the process of deciding which action to take. The degree of this belief is what we mean by "the probability that D caused E".[1]

Thus, a causality assessment method is a procedure for eliciting a "state of information" about a particular drug-event connection as input and delivering as output a "degree of belief" about the truth of the proposition that the drug caused the event to occur. In particular, the probability that D caused E is not an objective attribute of the facts of the case alone, but depends on the "state of information" of the person or persons carrying out the causality assessment.

Which, if any, causality assessment method should one use? The purpose of this paper is to address this question by developing criteria that causality assessment methods should satisfy and applying these criteria to evaluate currently available assessment methods. The main conclusion is that all current methods fail to satisfy at least some of these criteria.

The paper is organized as follows. Section 2 provides a critical review of the two criteria for causality assessment methods that can be found in previous papers on the subject. In section 3, we introduce and justify six criteria for causality assessment methods, and in section 4 we use these criteria to evaluate ten current methods.

2. Reproducibility and Validity: A Critical Review

The current literature on causality assessment focuses on two criteria for causality assessment methods: reproducibility and validity. Broadly speaking, a method is reproducible if it yields the same answer when applied more than once to the same case, and a method is valid if it yields the right answer. Stated thus, both criteria seem eminently reasonable, but we will argue that neither apply without substantial refinement to the causality assessment problem.

Reproducibility

Two different forms of the reproducibility criterion have been invoked in the causality assessment literature. Interobserver reproducibility requires different evaluators using the same causality assessment method on the same case to arrive at the same conclusion, while intraobserver reproducibility requires that a single evaluator be self-consistent when he uses

[1] An excellent discussion of probability as measure of degree of belief, and the use of probability in decision-making, can be found in [23].

the same method on the same case more than once. Only one paper [25] has examined the intraobserver reproducibility of a causality assessment method. In contrast, many papers have studied interobserver reproducibility for a variety of methods, starting with the important work of Karch et al. [15] and Koch-Weser, Sellers and Zacest [16], both of which showed the extent to which the unaided judgement of experts fails to attain interobserver reproducibility. Most standardized causality assessment methods do much better with respect to this criterion, and this fact has been the main argument advanced by their adherents in favor of the widespread adoption of these methods [3, 9, 22, 25].[2]

Evaluators with the same "state of information" about the connection between a drug D and an adverse event E should have the same "degree of belief" that D caused E. But different evaluators, even with the same access to such "objective" sources of information as individual case records and standard reference works, do not necessarily share the same "state of information" about the connection between D and E. For example, a case record might document that the onset of E occurred two days after administration of D, but the causal implications of this fact depend on a complicated web of theory and fact, including possible mechanisms by which D might cause E, the likelihood that an event of type E might occur from some other cause than D, and the timing distribution for onset of E for each possible D-causative mechanism and alternative causal agent. These theories and facts are all clouded by uncertainty, and each evaluator views them through the light of his own opinions and experience, whose relevance cannot be disregarded just because they are not "objective" and explicit.

Because of the complexity of the web of belief and experience that forms "states of information", we are inclined to doubt that different evaluators ought to share them. As a result, we believe that interobserver reproducibility is not a reasonable criterion for a causality assessment method. Rather, we think that it is of great importance that a method explicitly identify the key elements in an evaluator's "state of information", thus providing a possible mechanism for resolving differences between individual evaluators, and that it give each element its due weight in the causality assessment process. Specific criteria for achieving these goals are discussed in the next section of this paper.

To achieve interobserver reproducibility, the authors of most causality assessment methods tend to sweep uncertainty under the rug. They elicit only those aspects of the evaluator's "state of information" that are most public and objective, not necessarily the most important for distinguishing drug from nondrug causation. To avoid disagreements among evaluators, questions are posed with false precision: for example, the evaluator is typically asked to decide whether the timing of the event E is "consistent" with D-causation, avoiding the important issue of just <u>how</u> consistent the timing is, compared to alternative causative agents. Similarly, current methods elicit crude qualitative categorizations rather than quantitative estimates of the magnitude of effects; for example, in determining whether events of type E are known to occur as adverse reactions to D, most methods do not distinguish between incidence rates as disparate as 1 in 100 and 1 in 1000000. The tendency to ask questions that maximize the chance of agreement between evaluators leads to the

---

[2] A contrasting view is presented in [24].

elimination from consideration of those "entirely unexpected but vital imponderables which play a role in determining the value of a particular piece of evidence" [5], thus prejudicing M.N.G. Dukes against the use of any algorithmic causality assessment method. We share Dukes' concern, but we consider this tendency an argument not against standardized assessment methods, but against the criterion of interobserver reproducibility.

Validity

For certain kinds of probability assessment problems, it is possible to validate assessed probabilities by converting them into predictions about future observables. Suppose one wanted to evaluate a method for generating prognostic probability assessments -- for example, the method might produce an assessment that the probability for five-year survival given certain prognostic features is 1/2. Now if a cohort of patients with these features is assembled and if the probability assessment is correct, then in five years about 1/2 of them should have died. In this way, a collection of probability statements about future observables can be checked against experience to validate the method that produced the probabilities.

Unfortunately, probability assessments for events that have already occurred cannot be validated in this way. The aim of these assessments is not to predict what will happen but to gauge the probability of what did happen. Of course, if there were a pathognomonic test that could be applied retrospectively, the validity of a causality assessment method could be tested by seeing if a high proportion of those events judged "definite" turned out to be true adverse reactions, a smaller proportion of the "probables" and so forth, down to a negligible proportion of the "unrelateds". The problem is, of course, that such tests rarely exist. Even when they do, the result should be incorporated into the assessment method itself, rather than used as a test of validity in the test's absence.

Consequently, workers in the causality assessment field have turned to an alternative validation procedure, whose premise is: "to establish validity comparison with a standard is necessary" [25]. Such a procedure makes sense only if a reliable standard exists. The standard adopted in the causality assessment literature is the unaided judgement of experts, supplemented by some consensus-producing convention when the experts disagree with one another [7, 9, 14, 22, 25, 31, 30].[3] Now, to qualify as a "standard", a causality assessment must satisfy two requirements. First, the "state of information" on which it is based should be as complete and accurate as possible. Second, the technique used to convert this "state of information" into a "degree of belief" that the drug caused the event should merge the different elements in that "state of information" in a reasonable and unbiased way.

The unaided judgement of experts may achieve the first of these goals, but certainly fails to achieve the second. Cognitive psychologists have shown that the ability of the human brain to make unaided assessments of uncertainty in complicated situations is poor. We tend to scan the list of relevant facts, focus on one or two that seem particularly striking, and disregard the rest; and the situation is even worse when we try to argue from effect to cause,

[3] For more sophisticated attempts to achieve concordance with expert judgement, see [1] and [26].

as in the causality assessment problem.[4] Nor is the argument against the experts' unaided judgement as a standard for causality assessment merely theoretical: the adverse drug reaction literature provides ample evidence for its unreliability, documenting as it does the wide variability in assessments produced by different experts analyzing the same cases [see, for example 3, 9, 15, 16, 22, 25].

We believe that the effort to validate causality assessment methods by comparing their performance on particular cases to some "standard" method applied to those cases cannot succeed, because no true gold standard method exists. How then can we test the validity of these methods? Our answer is that criteria for validity must be applied not to the output of the methods, but to their procedures for merging and balancing the different elements in the "states of information" that serve as the method's input. That is, what must be validated is the logic of the methods, not their conclusions. In the next section, we shall turn to the task of developing criteria for the logic of causality assessment.

3. Six Criteria for Causality Assessment Methods

In this section, we present and justify six criteria for causality assessment methods. The first three criteria are intended to refine the reproducibility criterion discussed above, while the last three are designed to ensure that the methods process information in a logically valid way. The fourth criterion concerns the question of what information should be incorporable into the "state of information" elicited by a method, while the fifth and sixth have to do with how the evidence in the "state of information" should be evaluated and combined.

Criterion 1: Repeatability.

When the same "state of information" is used more than once as input, a causality assessment method should produce the same "degree of belief" as output.

Discussion: That a causality assessment method should satisfy this criterion is self-evident, but it is difficult to show that one does so, since it must first be established that two evaluators share a "state of information". Here are two procedures for checking that a method is repeatable, but neither is fully satisfactory, since the first is artificial and the second requires an unverifiable assumption. First, construct a set of artificial, fully explicit "states of information", each specifying the case and background information for a particular adverse clinical event; then have different subjects apply the causality assessment method, taking care that each use the given "state of information" as the sole input, and compare the corresponding outputs. Second, assume that the same individual is in the same "state of information" about the same case at two different time periods; this assumption may be approximately correct if the individual does not gain any relevant new knowledge or

[4] For a stimulating discussion of the psychology of reasoning in the face of uncertainty, see [12]; for a more detailed discussion of the difficulties of unaided causality assessments, see [21].

experience between the two times. With this assumption and in this situation, repeatability reduces to intraobserver reproducibility, which can be checked as in [25].

Criterion 2: Explicitness.

A causality assessment method should require that its user make explicit his "state of information", including the uncertainty he feels about each of its elements.

Discussion: This criterion imposes two requirements on a causality assessment method. The first concerns the content; the "state of information" must be elicited in as explicit form as possible. "States of information" are complicated; they include not only the objective facts of the case under consideration(such as the clinical condition and prior medical history of the patient, the symptoms and timing of the adverse event itself, and the nature and timing of the responses to dechallenge and rechallenge), but all the background knowledge and opinion that provide causal links between these facts and the various possible etiologies for the event. It is simply impossible to reason about complex and uncertain material in a coherent way, since it cannot all be held in mind simultaneously. Moreover, it is easier for different evaluators to pool their knowledge and assumptions and arrive at mutually agreeable assessments if they have each first laid out what they know and believe as explicitly as possible.

The second requirement of the explicitness criterion pertains to uncertainty. Evaluators should have to make explicit the degree of their uncertainty about each of the elements in their "states of information". This is necessary because many of the important elements simply cannot be known with certainty, and the effect that such elements have on the assessment should depend on the strength of the evaluators' belief in their truth.

How can one determine whether a causality assessment method satisfies the explicitness criterion? Clearly, the criterion is not an absolute standard; it is always a question of the degree to which it is satisfied. Nonetheless, it is possible to identify certain ways in which methods can fail to satisfy one or both requirements of this criterion.

A method may violate the first requirement by failing to deal explicitly with the contents of an evaluator's "state of information". It may do this by posing questions in such a way that specific elements of the "state of information" are left implicit or even entirely omitted. For example, the method may not require the user to specify alternative etiologies to the suspect drug (other drugs, drug interactions, the disease for which the suspect drug is administered, known comorbidity, or "unknown" etiology) or the type of reaction E might be to D (immunologic, dose-dependent, cytotoxic, and so forth). Even if these elements are elicited, however, the method may not make explicit the way in which they relate to one another and to the overall causality assessment. That is, the method may pose questions that call for the user to make implicit integrative judgements, rather than to take explicit account of the belief networks that connect the facts with their causal implications (for example, the user may be asked merely whether the timing of an event is "consistent" with drug causation, with no further clarification of the reasoning leading to such a global judgement).

A method may also fail to satisfy the requirement pertaining to the elicitation of the user's uncertainty. One way in which it may damp out uncertainty is to admit only "yes - no"

answers (or at most, to allow the user to respond "don't know", without making explicit the extent of his uncertainty).

Criterion 3: Explanatory Capability.

A causality assessment method must "explain" how it reaches its conclusions; that is, it must make it clear to the user why it produced the output "degree of belief" from the information it elicited.

Discussion: Causality assessment methods must make explicit the effect that each component of the elicited "state of information" has on the final assessment, and their users must be able to understand the rationale behind the transformation from information to degree of belief. Only if this process appears reasonable and not arbitrary can users gain confidence in the method, a confidence that is impossible if the method always functions as a "black box". Another advantage to explanatory capability is that it allows the user to identify what information turned out to be the most important in determining the overall causality assessment, and so he can focus his attention on the relatively few elements that drive the assessment, refining his opinions (and getting additional information) where it will make the most difference.

Criterion 4: Completeness.

A causality assessment method must respect R. A. Fisher's fundamental rule of uncertain inference: never throw information away.[5] That is, any fact, theory or opinion that can affect an evaluator's belief that a drug D caused an adverse event E must be incorporable by the method into the "state of information" on which the assessment is based.

This criterion has three particularly important corollaries:

> Corollary 1: Confronting uncertainty. Uncertain information cannot be ignored, nor can it be acted upon as though it were known with certainty.
>
> More specifically, suppose that a particular piece of information would affect the evaluator's degree of belief that D caused E, if the evaluator were certain that the information were true. Then a causality assessment method must take this information into account even if the evaluator is not certain about its truth, and the impact of this information on the causality assessment must depend on the extent to which the evaluator believes that the information is true.

---

[5] Fisher's expression of the rule is more elegant: "the logical characteristic, which has been too much overlooked, of all inferences involving uncertainty is that the rigorous specification of the nature and extent of the uncertainty by which they are qualified must in general involve the whole of the data, quantitative and qualitative, on which they are based." ([8], p. 113).

Corollary 2: Handling quantitative information quantatively. A causality assessment method must not require its users to interpret quantitative data qualitatively before accepting them as input, but must be able to elicit and process magnitudes directly.

Corollary 3: Using background information. When applied to a particular case of an adverse event E suspected of being a reaction to a drug D, a causality assessment method must take into account not only the details of that particular case, but all the background information relevant to the general connection between D and events of type E. This includes theories from the basic sciences, data from laboratory experiments and clinical experience, as well as epidemiological data about the relative incidence of events of type E when D is and is not administered to patients similar with respect to risk for events of type E to the particular patient under consideration.

Discussion: The completeness criterion prohibits a causality assessment method from excluding a priori any kind of information that an evaluator might find relevant to a particular assessment problem. The criterion can be applied to require that a method deal with information of a particular form, as in Corollaries 1 and 2, or with information with particular content, as in Corollary 3. To test whether the criterion is satisfied is not really possible, since to do so would require one to know in advance all possible sources and forms of information that might conceivably be relevant to some particular evaluator in some particular case. Nonetheless, it is possible to detect instances in which methods fail to satisfy the criterion, and this can be done, as the corollaries suggest, in some generality, without referring to the details of particular cases.

Corollary 2 may require special comment. As an example, suppose an epidemiological study carried out with subjects similar to the patient in the case under review reported an incidence rate of 1 event of type E per 1,000 exposures to drug D, compared with a rate of 1 event per 2,000 patients who did not receive D. The inferential content of this data may be difficult to assess, but it is surely different than it would be if the numbers were 1 per 1,000 compared to 1 per 10,000, say; or 1 per 10,000 compared to 1 per 100,000; moreover, it would hardly be sensible to select some arbitrary cutpoint and attempt to argue that if the ratio (or perhaps the difference?) of the two incidences exceeded the cutpoint then the meaning of the information was qualitatively different than if it did not. According to Corollary 2, a causality assessment method is deficient if it leaves to the user the task of extracting a qualitative summary of such numbers, instead of providing a method of deriving an appropriate contribution to the "degree of belief" directly from the numerical data itself.

Criterion 5 Etiological balancing.

It is not sufficient to evaluate case data just in terms of their concordance or discordance with the hypothesis that the drug D caused the event E. Rather, it is necessary to balance the likelihood of the data assuming that D caused E against the likelihood assuming alternative causes.

Discussion: According to this criterion, a causality assessment method can evaluate data concerning the clinical details of a particular case only by comparing the answers to two questions: how likely would this particular clinical picture be, if the event E really were an adverse reaction to the drug D, and how likely would it be, if the event E were unrelated to D? An alternative strategy for causality assessment, which we call the "D-causal hypothesis strategy", focuses only on the first of these questions. With this strategy, each piece of evidence is ranked as to how consistent it is with the "D-causal hypothesis" that D caused E (for example: if an adverse reaction to D of type E usually occurs between one and three days after initial exposure to D, and if in this case E happened two days after exposure, then this information about time of onset is highly consistent with the "D-causal hypothesis"). The case data is essentially divided into two piles: the data that are concordant with this hypothesis, and the data that are discordant with it. A method using the "D-causal hypothesis" strategy then delivers its "degree of belief" assessment as a function of how much higher the first pile is than the second (perhaps taking into account some cumulative measure of how concordant or discordant the elements in the piles are).

Methods using the "D-causal hypothesis" strategy need not ignore the problem of alternative etiologies. Typically, they consider the existence of obvious alternative etiologies for E (for example: other drugs, the condition for which the patient is being treated, or known comorbidity) as a particular piece of evidence that is discordant with the D-causal hypothesis. What they do not do is to consider how likely all the rest of the clinical evidence is, assuming that each of these alternatives were the true etiology for E. Nor do they deal with the more difficult questions of how likely it is that E has some other, unknown etiology, and how likely the clinical evidence is, assuming that some unknown cause is operating. These questions must be addressed if a method is to satisfy the etiological balancing criterion.

Why should the strategy of etiological balancing be preferred to the "D-causal hypothesis" strategy? The basic justification rests on the following premises: causality assessment is really a process of differential diagnosis; the evidentiary significance of each item of case information lies in its ability to distinguish between the "D-causal hypothesis" and alternative etiological hypotheses; and the way to measure the extent to which the information does distinguish between these hypotheses is to compare how much more (or less) likely the data are, given the hypothesis of D-causation than given the alternative hypothesis.

The "D-causal hypothesis" strategy and the strategy of etiological balancing are not equivalent; in many cases, they can yield different answers. Consider the following example. A patient is simultaneously started on dialysis and infusion with inferon. Within ten minutes, he experiences hypotension and shortness of breath. Is this adverse event a reaction to the dialysis or is it anaphylaxis induced by the inferon? From the point of view of the "D-causal" hypothesis, the timing of the event is highly concordant with inferon-causation, so this information should increase the "degree of belief" that the event is an anaphylactic reaction to inferon. On the other hand, the timing is just as consistent with dialysis-causation for the event, and so from the point of view of the strategy of etiological balancing, the information can neither increase nor decrease the "degree of belief" in D-causation.

Even more extreme examples can arise, especially when more than one drug could cause a particular event. In such cases, it is possible that data that is consistent with the hypothesis

that one of the drugs caused the event E may actually decrease the "degree of belief" that this drug caused E, because such data has even greater likelihood if another candidate drug is the cause.

Criterion 6: No a priori constraints on the effects of factors.

A causality assessment method should not limit a priori the effect that information about any particular factor (for example: background incidence, timing, response to dechallenge and rechallenge, patient's history) can have on the output "degree of belief".

Discussion: The following two examples illustrate the rationale behind this criterion, since they show that under certain conditions a positive response to dechallenge can determine the outcome of a causality assessment, while in other conditions the same kind of information can have virtually no effect on the assessment. Other examples, referring to timing, clinical characteristics of the event E, response to dechallenge and rechallenge, and background incidence are presented in appendix 1 at the end of this paper. As these examples show, methods can err in two directions: they can limit the magnitude of the effect of particular factors, or they can require that certain kinds of information (for example: positive response to dechallenge or rechallenge) always have an effect of greater magnitude than they may warrant in particular circumstances.

Example 1: A patient on hydralazine develops systemic manifestations of systemic lupus erythematosus, along with a positive anti-nuclear factor in the blood. All manifestations, including the anti-nuclear factor, clear without other treatment six months after discontinuing the hydralazine. Since systemic lupus caused by something other than the hydralazine would be very unlikely to clear spontaneously in this time period, the positive response to dechallenge in this case is very strong evidence of hydralazine causation. This example shows that the confirmatory effect of a positive response to dechallenge should not be limited a priori, while the next example shows that it is also might be necessary to discount the evidentiary significance of a positive response to dechallenge under different circumstances.

Example 2: A patient is hospitalized for cough, chills, fever and shortness of breath. While in the hospital, she is discovered to be hypertensive and methyldopa therapy is initiated. Within twenty four hours, she notes an increase in malaise and cough. Her serum amylase is then found to be elevated (760 units/dL, increasing to 1050 units/dL 12 hours later). Methyldopa is then discontinued (24 hours after detection of hyperamylasemia), and within 48 hours, serum amylase has fallen to 210 units/dL. This appears to be a positive response to dechallenge with methyldopa, but in fact the elevated amylase would be expected to drop within 24 hours regardless of etiology, and so the response to dechallenge is neutral with respect to the hypothesis of methyldopa causation.

4. Evaluation of Current Methodologies

In this section, we apply the criteria discussed in the previous section to evaluate ten standardized causality assessment methods. The methods evaluated are those of Karch and Lasagna [14] (denoted by KL); Dangomau, Evreux and Jouglard [4], in the version reported in [2] (denoted by DEJ); Kramer et al. [18] (denoted by KLHF); Blanc et al. [3] (B); Venulet, Ciucci and Berneker [31], with the scoring system reported in [29] (denoted by CG, for Ciba-Geigy); Emanueli and Saccheti [7] (ES); Naranjo et al. [25] (N); Jones [11], as reported in [28] (denoted by FDA); Lagier, Vincens and Castot [19] (LVC); and Stephens [27] (S).

For the purposes of this evaluation, we have divided the ten methods into three groups, as indicated in Table 1 below. The methods in the first group are quite short, posing fewer than ten questions to elicit states of information. The Group 2 methods are more complicated and elicit more complete states of information than those in the first group. LVC is the sole member of the third group. It is distinguished from the others in several respects, most strikingly by its use of etiological balancing.

TABLE 1 GOES HERE

We now turn to a detailed discussion of the performance of the methods with respect to each of the criteria (except criterion 1, repeatability, which we have not examined). The main conclusion of our evaluation is that none of these methods satisfy all the criteria; indeed, most of the methods violate most of the criteria.

Criterion 2: Explicitness

Recall that this criterion imposes two requirements; the first is that a method must elicit an explicit statement of its user's state of information. But states of information are too complicated to be elicited in response to ten "yes-no" questions, so Group 1 methods have to compromise with respect to this criterion. This compromise has two aspects: first, the elements of the state of information that provide the context for evaluating the evidentiary significance of the facts of the case are left implicit (for example: none ask the user to list all the alternative etiological candidates, or to consider the mechanism whereby D causes E, if E is an adverse reaction to D, and so forth); and second, the questions posed by the methods call for broad implicit integrative judgements from the user. Examples of such questions include

the following, from KL[6]: was there an "appropriate interval between agent-event?"; can the "event [be] reasonably explained by clinical state or other (nondrug) therapies?"; were the manifestations of the event "improved with dechallenge?". Clearly, the answers to such questions depend crucially on the subjective and implicit interpretation given to terms like "appropriate", "reasonably" and "improved". Some methods in this group avoid such integrative questions, but then only at the price of neglecting to obtain relevant information: for example, the only question posed by N about the timing of E is "did the adverse event appear after the suspected drug was administered?", which does not call for the user to decide about the "appropriateness" of the onset time with respect to D-causation, but neither does it allow any information about the length of time between administration of D and onset of E to affect the causality assessment.

In part because the user's uncertainty is already integrated into his answers to the questions they pose, Group 1 methods simply ignore the second requirement of the explicitness requirement, that the user make explicit the extent of his uncertainty about the information elicited from him. Only N even allows the user to respond "don't know", and N makes no further effort to elicit what the user believes, and to what extent, about the questions whose answers he "doesn't know" with certainty.

The methods in Group 2 elicit states of information much more explicitly than do those in Group 1. They require the user to provide many details about the case itself and the interpretative framework in which the case will be analyzed (alternative etiologies, mechanisms, etc. -- see, for example, the nine-fold classification of "category of adr" in CG or question 20 of the same method that asks the user whether E is a biologically plausible adverse reaction to D). In general, the Group 2 methods do well in eliciting the kind of detailed case information that can be operationally defined (see, for example, KLHF's dechallenge axis). On the other hand, they are not as successful in determining those elements of the state of information that depend on the user's uncertain beliefs and opinions. Unlike the methods in Group 1, Group 2 methods allow and even encourage[7] "don't know" answers, but they do not probe the user's degree of belief when he is not certain. Thus, an extra level of implicit judgement is involved in deciding whether to mark "U" for "unknown" (that is, not known with certainty) or to give a "certain" answer, to questions like the

---

[6] KL is not "worst case" by any means. Other examples include the following: "are there good alternative non-drug-related explanations for the event?" (DEJ: what does "good" mean in this context?). B asks the user to choose between the following three statements: "M1: The manifestations observed could not be due to the underlying disease. M2: They might be due to the disease, but the evolution was in favor of adverse reaction. M3: They could be due to the underlying disease and, moreover, the evolution was against the possibility of an adverse reaction." FDA asks "does event have a reasonable temporal association with use of the drug?" and "could the event be due to an existing clinical condition?", and only allows yes-no answers (no "don't know" or "ambiguous").

[7] "We insist that [the assessment] reflects information, not assumptions. That means that if certain information is not given, it should not be assumed "no", but it should be marked as unknown." ([29], p. 317), discussing the use of CG).

following (from CG): "Drug/ADR interval compatible with the event? (Possible answers: typical; compatible; incompatible; or U)"; "ADR occurrence facilitated by the disease treated or by concomitant diseases? (Possible answers: yes; no; or U)"; "Other contributory factors (habits, environment etc.)? (Possible answers: yes; no; or U)".[8] How can a user's information relative to such questions be elicited properly if his uncertainty about the data and his interpretative framework is not explicitly assessed?

LVC works completely differently from the other methods. For each of nine factors, the user is presented with a list of general considerations and then asked to measure his uncertainty about D-causation, with respect to the evidence relevant to the factor in question, on an uncalibrated scale of 0 to 1. These scores are then merged into three more general scores on the same scale (chronological, symptomological and relative risk), and then these are merged once more into an overall judgement of "imputation", that also takes into account the user's assessment of the completeness and reliability of the evidence. Each of these successive scorings is integrative and implicit. Thus, this method fails to elicit the state of information explicitly; in fact it makes no effort to do so at all. On the other hand, its purpose is precisely to obtain an explicit measurement of the user's uncertainty about the information available to him and its implications for D-causation.

Criterion 3: Explanatory Capability

All the methods except LVC determine the output degree of belief as explicit functions of the answers provided to the questions they pose. Thus, it is possible for users of these methods to trace the effect of any particular answer on the overall assessment. On the other hand, the more questions there are or the more complicated the scoring functions, the more difficult it is to sort out the effects of general assumptions or opinions (for example, about the mechanism of the adverse reaction or about the possible role of the clinical condition in facilitating or causing E), since these typically effect the answers to more than one question. Group 2 methods are particularly confusing in this regard; KLHF because the number of questions it poses is large, and CG and S because their scoring functions are especially complicated, the points attached to each question varying with the question and with the category of adverse reaction (nine in CG, two in S) selected by the user.

However, explanatory capability requires more than making clear to the user how an assessment was reached; the method must make clear why it was reached as well. No method achieves this goal. In all of them the scoring attached to each question appears arbitrary and is not explicitly justfied, and the same is true for the procedure by which these scores are combined into an overall assessment (see the discussion of criterion 6 below for details on one particular aspect of this arbitrariness). For example, KLHF in effect evaluates the absence of dechallenge as evidence in favor of D-causation by awarding a point on the dechallenge axis if no dechallenge occurs and the sum of the scores on the first two axes

[8] CG is equipped with a five page "List of Definitions and Clarifications" to help the user answer its questions, but there is no comment about the degree of certainty the user must feel before he should give an answer to these (and similar) questions different from U.

exceeds two; similarly, DEJ gives a score of three to an irreversible event with "consistent timing", while scoring only two for a reversible event with equally "consistent timing" for which the response to dechallenge is "not assessable", either because dechallenge was not attempted or because the symptoms of the event were treated directly. Even when the scoring is not as obviously arbitrary as in these examples, the reasoning behind it is neither explained nor apparent.

LVC is again an exceptional case. Its authors take the point of view that the assessment process cannot be reduced to a calculation that provides its own justification: "imputation is neither a sum nor a product of scores; rather, it is an assessment that cannot be given by a well-defined mathematical function" [19, p. 311]. As such, they make no attempt to satisfy either of the requirements of this criterion.

Criterion 4: Completeness

Group 2 methods elicit a good deal more information than group 1 methods do; CG and KLHF are particularly wide-ranging and thorough in the case data they explicitly take into account, and S and LVC cover the same or more ground in a less explicit way. The group 1 methods do not process detailed information about any factor, and they completely exclude whole categories of potentially relevant information: DEJ, B, KL and FDA do not inquire about previous experience of the patient with D or similar drugs or events of type E, and KL, FDA and ES cannot take into account details of E such as laboratory findings or serum levels in assessing the probability of D-causation.

None of the methods except LVC attempts to elicit information about the prior incidence of events of type E with and without D involvement, except in the crudest, most qualitative way. FDA ignores information about previous D-E connections entirely, while N's questions 1 and 5 -- "are there previous conclusive reports on this reaction?" and "are there alternative causes (other than the drug) that could on their own have caused the reaction?" -- are typical of the other Group 1 method's treatment of prior incidence. The Group 2 methods refine N's approach, for example by considering what the alternative causes might be (and hence how many different etiological contenders there are). Nonetheless, the questions dealing with the incidence of events of type E from the various possible causes still call for broad categorical answers; for example, CG asks if "adverse event of rare spontaneous occurrence?", and S takes account of how long D has been on the market relative to the number of reported D-E connections to determine to which of 12 "incidence" classes the adverse reaction belongs.

The failure of the methods to take into account more refined prior incidence estimates is related to their general position on quantification. None except LVC process information in quantitative form, and none except LVC deliver their output degrees of belief as numbers. Unfortunately, although LVC deals with numbers, it does not use mathematics to process them and does not define their meanings, but uses the interval from 0 to 1 purely as an analogue scale. Thus, Corollary 2 of the completeness criterion is uniformly violated by the methods under review.

Criterion 5: Etiological balancing

With respect to this criterion, there is no difference between Group 1 and Group 2 methods: they all use the D-causal hypothesis approach to processing case information and thus violate the criterion. The role that alternative etiologies play in their assessments varies somewhat, however. On one extreme is ES, who ask if "the possibility that clinical state or therapies may explain the event can be ruled out?", and if the answer is "no", it is impossible to attain an assessment more favorable to D-causation than "doubtful". In contrast, FDA only takes account of alternative etiologies if there is a positive response to dechallenge and no rechallenge, in which case if the event "could be due to an existing clinical condition", the rating is "possible", otherwise "probable". Neither these methods nor N, KL, DEJ, S or CG takes any account of how likely the data is given nondrug causation for data about timing, characteristics of the event, or rechallenge; S and CG do consider whether a specific treatment for E was administered when evaluating dechallenge data, but they do not consider how likely it is for E to abate if it is not treated and is not an adverse reaction to D. KLHF consider this last possibility in their dechallenge axis, which elicits and analyzes data as consistently with etiological balancing as any method except LVC; but the alternative candidates, timing and rechallenge axes of KLHF considers only the existence of other etiologies, not the relative likelihood of the data given these compared to D-causation.

B has three criteria, and the ratings with respect to the first two, time sequence and response pattern, are strictly of the D-causal hypothesis type. The third criteria is "role of underlying disease(s)", and the user is asked to decide whether "the manifestations observed could not be due to the underlying disease," or "they might be due to the disease, but the evolution was in favor of adverse reaction," or "they could be due to the underlying disease and, moreover, the evolution was against the possibility of an adverse reaction." This choice clearly requires etiological balancing, but at a global level, so that the user essentially is required to carry out a complete causality assessment in his head to decide among the three alternatives; to carry out this assessment, he must surely integrate timing and response pattern information, so it is not clear what additional role the first two criteria play.

In contrast to the other methods, LVC explicitly adopts the etiological balancing strategy. The point of each of the assessments required in this method is to balance the likelihood of specific information given D-causation against its likelihood given alternative etiologies. A score of 0 or 1 corresponds to certainty, with respect to alternative and D-causation respectively, and a score of 0.5 corresponds to neutrality (or equally favoring evidence) with respect to the two hypotheses.

Criterion 6: No a priori constraints on factor effects

All the methods, except LVC, place constraints on factor effects. In general, these constraints are severe and quite idiosyncratic, and they lead to unreasonable assessments when the methods are applied to cases with features similar to those sketched in the discussion of this criterion in Section 3 and Appendix 1. S and CG allow the most flexibility in the extent to which information about different factors can affect the final assessment, but it is difficult to judge CG in this respect because its complete scoring system has not been

published. LVC gives no procedures for combining the scores on its nine factors, so its performance with respect to this criterion cannot be evaluated.

Here are some examples of the kinds of constraints inherent in the methods. ES gives a maximum rating of "doubtful" unless "the possibility that clinical state or therapies may explain the event can be ruled out," and unless an adverse reaction is known, the highest rating is "possible" if rechallenge has not been attempted. If the timing relation between D and E was not "appropriate", then D and E are judged "unrelated"; otherwise, information about timing plays no role at all in the assessment (even if E is an allergic reaction occurring immediately after administration of D at the site of administration!). Also, a positive response to dechallenge must produce an assessment of "almost definite" (compare to example 2 in Section 3), while a negative response to rechallenge cannot reduce the chance of D-causation.

These examples from ES are not at all atypical; a similar list could be developed for any of the other methods, except possibly S and CG and, of course, LVC. Constraints can be positive or negative: in the examples above, the constraint on timing is negative, while the constraint on positive response to dechallenge is positive (it is forced to bear more weight than it may rightfully carry in special circumstances). Some of the methods mainly impose negative constraints; this is true, for example, for N and KLHF, which have additive scoring systems with a maximum of 2 points per factor (generally just one), with 10 and six factors respectively. Thus, KLHF cannot produce a "definite" assessment, regardless of the strength of the evidence about timing and prior frequencies, unless a positive response to dechallenge is observed.

S and CG do better with respect to this criterion, because they allow strong evidence in various factors to produce very high scores for that factor, and S in particular avoids negative constraints (see for example factor 4, rechallenge, where both "difficult to judge due to changes in underlying disease -- no response" and "difficult to judge due to changes in underlying disease -- some response", a weak negative and weak positive response respectively, score a neutral 5 and 6 respectively). But they are not completely free of constraints either: the scores CG assigns to dechallenge and rechallenge depend just on whether the response was positive or negative, and not on special circumstances (assuming the event is reversible and is not treated directly); in particular, in the "dose-dependent" scoring category, negative response to rechallenge is always taken as strong evidence against D-causation (-25) and positive response strong evidence for D-causation (+30), in contrast to example 8 above.

Conclusions

None of the methods performs satisfactorily when measured against our criteria. More specifically, here is a summary of the most important results from our evaluation:

1) Only one of the methods (LVC) employs the strategy of etiological balancing to analyze case information. All the rest use the D-causal hypothesis strategy, with at most minor adjustments.

2) None of the methods tries to provide an explicit rationale for the way it converts a state of information into a degree of belief. In particular, the methods combine the scores assigned to the various factors in essentially arbitrary ways, typically with strong a priori constraints imposed on the effect that information about each factor can have on the overall assessment.

3) No method has satisfactory procedures for incorporating quantitative information or measuring the user's uncertainty about the information elicited from him.

4) The methods that elicit the most complete states of information tend to present the greatest difficulty in tracing the effects that particular pieces of input have on the output. On the other hand, the methods for which it is easiest to trace these effects either fail to ask about important sources of information or call for substantial implicit integrative judgements on the part of their users.

In conclusion, if the criteria are reasonable, the methods are not. Are there alternatives? Certainly one exists: global introspection [21], the unaided judgement of experts. But it can be judged according to these criteria as well, and its explicitness and explanatory capability are clearly poor, while it seems unlikely that it satisfies the criteria of completeness and no a priori constraints. In our opinion, a new approach to causality assessment is required. In collaboration with Judith Jones, Claudio Naranjo and Michael Kramer, we plan to present such an approach in a future article and show that it conforms to the criteria developed here.

REFERENCES

1. Bastin, C., P. Wertheimer and J. Venulet (1984). Adverse drug reactions -- a computer-assisted application of correspondence analysis for automatic causality assessment. Meth. Inform. Med 23, 183-188.

2. Begaud, B. (1984). Standardized assessment of adverse drug reactions: the method used in France. Drug Information Journal 18, 275-282.

3. Blanc, S., P. Leunberger, J.P. Berger, E.M. Brooke and J.L. Schelling (1979). Judgements of trained observers on adverse drug reactions. Clinical Pharmacology and Therapeutics 25, 493-498.

4. Dangoumau, J., J.C. Evreux and J. Jouglard (1978). Methode d'imputabilite des effets adverses des medicaments. Therapie 33, 373-381.

5. Dukes, M.N.G. (1984). Adverse reactions: a changing challenge. Proceedings of the Second World Conference on Clinical Pharmacology and Therapeutics (American Society for Pharmacology and Experimental Therapeutics, Bethesda, Maryland), 223-232.

6. Emanueli, A. and G. Sacchetti (1982). Postmarketing surveillance methodology as applied in a pharmaceutical medical department. Drug Safety -- Progress and Controversies (Auriche, Burke and Duchier, eds., Paris, Pergamon Press), 265-273.

7. Emanueli, A. and G. Sacchetti (1980). An algorithm for the classification of untoward events in large scale clinical trials. Agents and Actions 7, 318-322.

8. Fisher, R.A. (1973). Statistical Methods and Scientific Inference (Third Edition, New York, Hafner).

9. Hutchinson, T.A., J.M. Leventhal, M.S. Kramer, F. Karch, A. Lipman and A.R. Feinstein (1979). An algorithm for the operational assessment of adverse drug reactions. II Demonstration of reproducibility and validity. JAMA 242, 633-638.

10. Irey, N.S. (1976). Tissue reactions to drugs. American Journal of Pathology 82, 617-647.

11. Jones, J. (1982). Adverse drug reactions in the community health setting: Approaches to recognizing, counseling and reporting. Family and Community Health 5, 58-67.

12. Kahneman, D., P. Slovic and A. Tversky (1982). Judgement under Uncertainty: Heuristics and Biases (New York, Cambridge University Press).

13. Karch, F. and L. Lasagna (1975) Adverse drug reactions -- a critical review. JAMA 234, 1236-1241.

14. Karch, F. and L. Lasagna (1977). Toward the operational identification of adverse drug reaction. Clinical Pharmacology and Therapeutics 21, 247-254.

15. Karch, F., C. Smith, B. Kerzner, J. Mazullo, M. Weintraub and L. Lasagna (1976). Adverse drug reactions -- a matter of opinion. Clinical Pharmacology and Therapeutics 19, 489-492.

16. Koch-Weser, J., E.M. Sellers and R. Zacest (1977). The ambiguity of adverse drug reactions. European Journal of Clinical Pharmacology 11, 75-78.

17. Kramer, M.S (1981). Difficulties in assessing the adverse effects of drugs. British Journal of Clinical Pharmacology 11, 105S-110S.

18. Kramer, M.S., J.M. Leventhal, T.A. Hutchinson and A.R. Feinstein (1979). An algorithm for the operational assessment of adverse drug reactions: I. Background, description, and instructions for use. JAMA 242, 623-633.

19. Lagier, G., M. Vincens and A. Castot (1983). Imputabilite en pharmacovigilance: Principes de la methode appreciative ponderee et principales erreurs a eviter. Therapie 38, 303-318.

20. Lagier, G., M. Vincens, A. Castot and M.L. Efthymiou (1982). Diagnostic differential des leucopenies medicamenteuses. Therapie 37, 377-384.

21. Lane, D. (1984). A probabilist's view of causality assessment. Drug Information Journal 18, 323-330.

22. Leventhal, J.M., T.A. Hutchinson, M.S. Kramer and A.R. Feinstein (1979). An algorithm for the operational assessment of adverse drug reactions. III Results of tests among clinicians. JAMA 242, 1991-1994.

23. Lindley, D.V. (1971). Making Decisions (Wiley, New York).

24. Louik, C., P. Lacouture, A. Mitchell, R. Kaufmann, F. Lovejoy, S. Yaffe and S. Shapiro (1985). A study of adverse reaction algorithms in a drug surveillance program. Clinical Pharmacology and Therapeutics 38, 183-187.

25. Naranjo, C., U. Busto, E.M. Sellers, P. Sandor, I. Ruiz, E.A. Roberts, E. Janecek, C. Domecq and D.G. Greenblatt (1981). A method for estimating the probability of adverse drug reactions. Clinical Pharmacology and Therapeutics 30, 239-245.

26. Salomon, R. and P. Peytour (1981). Recherche du poids des criteres dans un algorithme d'imputabilite. Therapie 36, 229-231.

27. Stephens, M. (1984). Assessment of causality in industrial settings. Drug Information Journal 18, 307-314.

28. Turner, W. (1984). The Food and Drug Administration algorithm. Drug Information Journal 18, 259-266.

29. Venulet, J. (1984). The Ciba-Geigy approach to causality. Drug Information Journal 18, 315-318.

30. Venulet, J. and A. Ciucci (1982). Assessment of the cause and effect relationship of suspected adverse drug reactions. Drug Safety -- Progress and Controversies (Auriche, Burke and Duchier, eds., Paris, Pergamon Press), 83-91.

31. Venulet, J., A. Ciucci and G.C. Berneker (1980). Standardized assessment of drug-adverse reaction associations -- rationale and experience. International Journal of Clinical Pharmacology, Therapy and Toxicology 18, 381-388.

32. Venulet, J., G.C. Berneker and A. Ciucci (eds) (1982). Assessing Causes of Adverse Drug Reactions (London, Academic Press).

TABLE 1: THE METHODS

Method	Authors	References	Group
B	Blanc et al.	[3]	1
CG	Venulet et al.	[29], [30], [31]	2
DEJ	Dangoumau et al.	[4], [2]	1
ES	Emanueli and Saccheti	[6], [7]	1
FDA	Jones et al.	[11], [28]	1
KL	Karch and Lasagna	[14]	1
KLHF	Kramer et al.	[18], [9], [22]	2
LVC	Lagier et al.	[19], [20]	3
N	Naranjo et al.	[25]	1
S	Stephens	[27]	2

Note: Group 1 methods are the shortest, with fewer than ten questions; Group 2 and 3 methods are more complicated.

APPENDIX 1: ADDITIONAL EXAMPLES RELATED TO CRITERION 6

Timing

Example 1: A patient with clinically stable lupus erythematosus develops renal dysfunction (an increase in serum creatinine from 1 to 2 mg%) one day after starting aspirin therapy. The renal dysfunction could conceivably be due to the underlying disease, but if so it would have approximately the same chance of appearing at any stage of this life-long disease. Hence, in this case the concordance of the timing with the hypothesis of aspirin-causation, compared to the diffuseness of the timing distribution for the alternative etiology, gives fairly conclusive evidence for drug causation, and so methods should not bound the potential effect of timing information from above.

Example 2: As above, except that the serum creatinine increases one hour after the first aspirin tablet is taken. In this case, the latency period is too short (even with renal shut-down it would take at least 12 hours for the serum creatinine to rise by 1 mg%), and so information about timing conclusively refutes drug causation, showing that the effect of timing information on probability of D-causation should not be bounded from below, even when the patient received D before E occurred.

Clinical characteristics

Example 3: A child in an intensive care unit unexpectedly develops cardiac arrest and dies while receiving digoxin therapy. Blood tests show the digoxin level is twenty times the therapeutic level;the assay is repeated, with the same result. This information makes it much more likely that digoxin was the cause of death. Thus, the effects of information about serum levels (or other clinical characteristics associated with the event E) should not be bounded from above.

Dechallenge

Example 4: The examples presented in section 3 showed the effect of positive response to dechallenge could vary considerably depending on circumstances. Similiarly, negative response to dechallenge can provide conclusive evidence against drug causation or it can be neutral between drug and nondrug etiologies. This example illustrates the former possibility, the next the latter. A patient develops symptoms of dizziness associated with postural hypotension after beginning alpha-methyldopa therapy for hypertension. These symptoms persist for at least one year after the medication is stopped. Clearly, methyldopa causation is highly unlikely.

Example 5: A patient taking an estrogen-containing contraceptive suffers a stroke. She is taken off the pill, with no effect on recovery of function from the stroke. In this case, the negative response to dechallenge does not argue against estrogen causation for the stroke.

Rechallenge

Example 6: Positive response to rechallenge often provides conclusive evidence of drug causation. The following example shows that it need not do so. A female adolescent who has recently become sexually active develops non-specific vaginitis shortly after starting to take oral contraceptives. By the time she sees a physician for her vaginal discharge she has broken off her relationship with her previous boyfriend. The physician treats her vaginitis and discontinues the oral contraceptive. A year later, she develops a new relationship, restarts the contraceptive and again develops non-specific vaginitis. In this case, it is impossible to distinguish the possible causal roles of sexual activity and the oral contraceptive, since their effects are mutually confounded. The recurrence with rechallenge does not increase the probability of drug causality, since the sexual activity "recurred" simultaneously.

Background incidence

Example 7: In some cases, the most important information differentiating between the various etiological candidates is simply background incidence, not data concerning the particular case at hand. For example, suppose a patient who has been undergoing dialysis for some time with no previous adverse effects receives both diphenhydramine and iron dextran in dialysis for the first time and suffers acute anaphylaxis. Both drugs are possible causes, but iron dextran is associated much more frequently with allergic reactions than is diphenhydramine, and so the anaphylaxis is much more likely to be caused by the iron dextran than the diphenhydramine. Thus, no method should bound a priori the effect of background incidence on the probability of drug causation.

Example 8: In contrast to example 9, here is a case in which background incidence strongly favors nondrug causation. A patient in septic shock, severe pulmonary edema and renal shutdown is given hemodialysis to remove fluid as a final effort to save his life. The patient becomes increasingly hypoxic and hypotensive and dies within an hour of beginning dialysis. Although hemodialysis can produce these effects, they were sufficiently likely as the natural course of the patient's disease that the probability that the reactions were due to the hemodialysis is very small.